PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-066196
(43)Date of publication of application : 09.03.1999

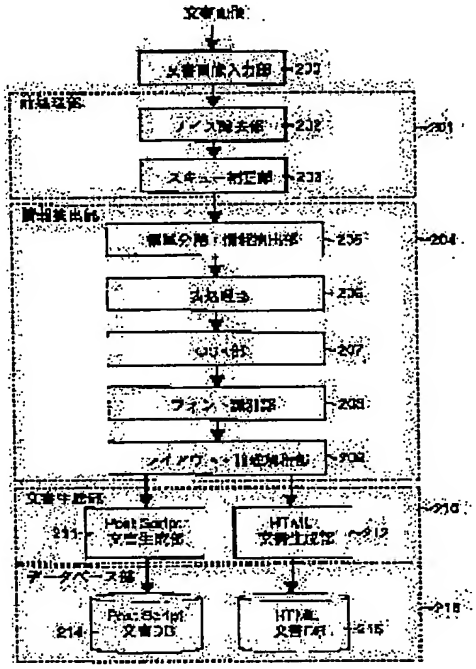
(51)Int.Cl. G06F 19/00
G06F 17/21
G06T 7/00
G06K 9/20
H04N 1/21
H04N 1/40

(21)Application number : 09-220426
(22)Date of filing : 15.08.1997

(71)Applicant : RICOH CO LTD
(72)Inventor : SAITO TAKASHI
ABE TEI
KOUCHI TSUKASA

(54) DOCUMENT IMAGE RECOGNITION DEVICE AND COMPUTER-READABLE RECORDING MEDIUM WHERE PROGRAM ALLOWING COMPUTER TO FUNCTION AS SAME DEVICE IS RECORDED

(57)Abstract:
PROBLEM TO BE SOLVED: To generate documents in various format complying with usages such as a document whose reproduction is given priority and a document whose contents are made important.
SOLUTION: This device is equipped with a document image input part 200 which inputs a document image generated by optically reading a paper document, a preprocess part 201 which performs noise removal and a skew correcting process for the inputted document image, an information extraction part 204 which performs a recognizing and extracting process for a character area including character strings and/or a an image area including images of a graph, a table, a photograph or the like, and a character recognizing process for the character strings in the extracted character area and also analyzes the layout of the document image to extract layout information, a document generation part 210 which generates a PostScript document and an HTML document according to the character recognition result and layout information extraction result, and a data base part 213 which stores the generated PostScript document and HTML document respectively.



LEGAL STATUS
[Date of request for examination] 13.06.2002
[Date of sending the examiner's decision of rejection] 04.02.2003
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]

(19) 日本国特許庁(JP)　(12) 公開特許公報（A）　(11) 特許出願公開番号

特開平11－66196

(43) 公開日　平成11年(1999) 3月9日

(51) Int. Cl.⁶		識別記号	F I		
G06F	19/00		G06F	15/22	G
	17/21		G06K	9/20	340 C
G06T	7/00		H04N	1/21	
G06K	9/20	340	G06F	15/20	538 A
H04N	1/21			15/70	330 Q

審査請求　未請求、請求項の数4　OL　(全16頁) 最終頁に続く

(21) 出願番号　特願平9-220426

(22) 出願日　平成9年(1997) 8月15日

(71) 出願人　000006747
株式会社リコー
東京都大田区中馬込1丁目3番6号

(72) 発明者　齋藤　高志
東京都大田区中馬込1丁目3番6号　株式会社リコー内

(72) 発明者　阿部　悌
東京都大田区中馬込1丁目3番6号　株式会社リコー内

(72) 発明者　幸地　司
東京都大田区中馬込1丁目3番6号　株式会社リコー内

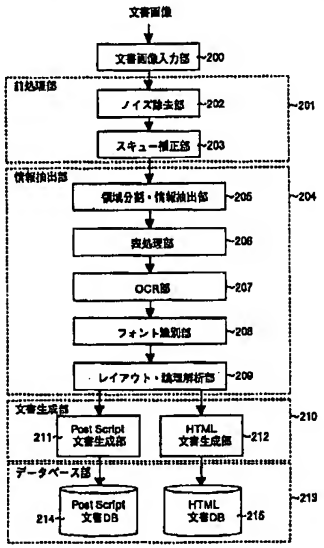
(74) 代理人　弁理士　酒井　宏明

(54) 【発明の名称】文書画像認識装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体

(57) 【要約】

【課題】　紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態の文書を生成すること。

【解決手段】　紙文書を光学的に読み取ることによって生成した文書画像を入力する文書画像入力部２００と、入力した文書画像に対してノイズ除去およびスキュー補正処理を行う前処理部２０１と、文書画像から文字列を含む文字領域および／または図，表，写真等の画像を含む画像領域の認識・抽出処理，抽出した文字領域の文字列についての文字認識処理、および文書画像のレイアウトを解析し、レイアウト情報の抽出処理を行う情報抽出部２０４と、文字認識結果およびレイアウト情報抽出結果に基づいて、ＰｏｓｔＳｃｒｉｐｔ文書およびＨＴＭＬ文書を生成する文書生成部２１０と、生成したＰｏｓｔＳｃｒｉｐｔ文書およびＨＴＭＬ文書をそれぞれ格納するデータベース部２１３と、を備えている。



1

【特許請求の範囲】

【請求項１】　紙文書を光学的に読み取ることによって生成した文書画像を入力する入力手段と、

前記入力手段を介して入力した文書画像から文字列を含む文字領域および／または図，表，写真等の画像を含む画像領域を認識して抽出する領域抽出手段と、

前記領域認識手段で抽出した文字領域の文字列について文字認識処理を行う文字認識手段と、

前記領域抽出手段の抽出結果に基づいて、前記文書画像のレイアウトを解析し、レイアウト情報を抽出するレイアウト情報抽出手段と、

前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、ページ記述言語を用いた第１の文書を生成する第１の文書生成手段と、

前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、構造化記述言語を用いた第２の文書を生成する第２の文書生成手段と、

前記第１および第２の文書生成手段で生成した第１および第２の文書をそれぞれ格納する格納手段と、

を備えたことを特徴とする文書画像認識装置。

【請求項２】　前記第１の文書は、ＰｏｓｔＳｃｒｉｐｔ形式またはＰＤＦ形式によって表現された文書であることを特徴とする請求項１に記載の文書画像認識装置。

【請求項３】　前記第２の文書は、ＳＧＭＬ，ＨＴＭＬまたはＸＭＬによって表現された文書であることを特徴とする請求項１に記載の文書画像認識装置。

【請求項４】　前記請求項１～３のいずれか１つに記載の文書画像認識装置の各手段としてコンピュータを機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、紙文書を光学的に読み取ることによって得た文書画像から文字コードを抽出するという単なる文字認識処理を行うだけではなく、紙文書の持つ様々な情報を抽出して利用することにより、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を生成することができるようにした文書画像認識装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録に関する。

【０００２】

【従来の技術】パーソナルコンピュータやネットワークの急速な普及は、オフィスにおける文書の電子化を促進することとなった。ところが、紙は閲覧性が良い等の利便性があるため、紙文書の生産は未だ止むことがなく、紙の形で保存・流通されている文書は未だ多量に存在しているのが現状であり、このギャップがオフィスワークにおける生産効率の低下の一因となっている。

2

【０００３】つまり、文書の作成，流通，閲覧，管理，再利用といった一連の流れにおいて、文書は紙に記録された形とコンピュータ上のデータという形で存在しており、このように文書が紙に記録された形とコンピュータ上のデータという形で存在しているのは、相互の変換コストが高いということが原因となっている。

【０００４】上記問題を解決する手段の一つとして、ＯＣＲ（Ｏｐｔｉｃａｌ　Ｃｈａｒａｃｔｅｒ　Ｒｅａｄｅｒ）を挙げることができる。このＯＣＲを用いることにより、紙文書をスキャナ等で読み取って文書画像を生成した後、文書画像中の文字列を文字コードに変換することができる。

【０００５】

【発明が解決しようとする課題】しかしながら、上記ＯＣＲは、文書画像中の文字列を文字コード情報に変換することは可能であるが、元の紙文書のレイアウト等まで抽出することができないため、ＯＣＲによって生成された文書は単なるテキストの羅列であって、元の文書画像が持つ様々な情報を十分に利用することができないという問題点があった。換言すれば、文書は、作成者の意図に応じて様々なレイアウト処理が施されているが、単にＯＣＲを用いただけでは、文書のレイアウトを抽出することができず、元の文書のレイアウトを再現した文書を新たに生成したり、元の文書のレイアウトを利用して新たなレイアウトの文書を生成したりすることはできなかった。

【０００６】本発明は上記に鑑みてなされたものであって、紙文書を光学的に読み取ることによって得た文書画像から文字コードを抽出するという単なる文字認識処理を行うだけではなく、紙文書の持つ様々な情報を抽出して利用することにより、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を生成することができるようにすることを目的とする。

【０００７】

【課題を解決するための手段】上記目的を達成するため、請求項１の文書画像認識装置は、紙文書を光学的に読み取ることによって生成した文書画像を入力する入力手段と、前記入力手段を介して入力した文書画像から文字列を含む文字領域および／または図，表，写真等の画像を含む画像領域を認識して抽出する領域抽出手段と、前記領域認識手段で抽出した文字領域の文字列について文字認識処理を行う文字認識手段と、前記領域抽出手段の抽出結果に基づいて、前記文書画像のレイアウトを解析し、レイアウト情報を抽出するレイアウト情報抽出手段と、前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、ページ記述言語を用いた第１の文書を生成する

第1の文書生成手段と、前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、構造化記述言語を用いた第2の文書を生成する第2の文書生成手段と、前記第1および第2の文書生成手段で生成した第1および第2の文書をそれぞれ格納する格納手段と、を備えたものである。

【0008】また、請求項2の文書画像認識装置は、請求項1に記載の文書画像認識装置において、前記第1の文書が、PostScript形式またはPDF形式によって表現された文書であるものである。

【0009】また、請求項3の文書画像認識装置は、請求項1に記載の文書画像認識装置において、前記第2の文書が、SGML，HTMLまたはXMLによって表現された文書であるものである。

【0010】さらに、請求項4のコンピュータ読み取り可能な記録媒体は、前記請求項1～3のいずれか1つに記載の文書画像認識装置の各手段としてコンピュータを機能させるためのプログラムを記録したものである。

【0011】

【発明の実施の形態】以下、本発明の文書画像認識装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体の実施の形態について、添付の図面を参照しつつ詳細に説明する。

【0012】〔実施の形態1〕実施の形態1の文書画像認識装置は、紙文書を光学的に読み取って生成した文書画像を入力し、入力した文書画像に基づいて、コンピュータ上で利用可能な文書を生成するものである。換言すれば、オフィスワーク等においては、文書を生産し、流通し、閲覧し、管理し、利用するという一連の流れがあるため、実施の形態1の文書画像認識装置は、上記流通および閲覧を考慮した文書と、利用を考慮した文書を生成することができるようにしたものである。

【0013】ここで、流通および閲覧を考慮した文書とは、文書作成者の意図をできるだけ正確に伝えることができるように、オリジナルの紙文書が持つレイアウト情報まで忠実に再現することを目的とした文書のことである（紙文書のメタファとしてのWYSIWYGを保存した形の文書）。紙文書をコンピュータ上の文書で忠実に再現することができるようにするために、実施の形態1の文書画像認識装置は、文字や画像を統一的に記述することができるページ記述言語を用いて文書を生成するものである。なお、実施の形態1の文書画像認識装置においては、ページ記述言語として、PostScriptを用いることにする。以下では、このPostScriptを用いて生成した文書のことをPostScript文書と定義することにする。また、PostScriptに代えて、PDF（Portable Document Format：アドビシステム社が公開しているファイル形式）形式を用いることもできる。

【0014】また、利用を考慮した文書とは、オリジナルの紙文書が持つレイアウトにとらわれず、文書の内容を優先したコンピュータならではの形を持った文書のことである。このように、レイアウトよりも内容を優先した文書を生成するために、実施の形態1の文書画像認識装置は、SGML（Standard Generalized Markup Langage）やHTML（HypertextMarkup Langage）、XML（Extensible Markup Langage）等の構造化記述言語で表現することによって文書を生成し、文字や画像が混在した紙文書をハイパーテキスト化するものである。なお、実施の形態1の文書画像認識装置においては、構造化記述言語として、HTMLを用いることにする。以下では、このHTMLを用いて生成した文書のことをHTML文書と定義することにする。

【0015】図1は、実施の形態1の文書画像認識装置のシステム構成を示す構成図である。図1に示す文書画像認識装置は、紙文書を光学的に読み取って文書画像を生成するカラースキャナ100，モノクロスキャナ101およびネットワークスキャナ102と、ファクシミリ装置103から送信された文書画像を受信するファックスモデム104と、カラースキャナ100，モノクロスキャナ101，ネットワークスキャナ102およびファックスモデム104から文書画像を入力し、入力した文書画像に基づいて、PostScript文書およびHTML文書を生成する文書画像処理サーバ105aと、紙文書を光学的に読み取って文書画像を生成するディジタル複合機106と、ディジタル複合機106から文書画像を入力し、入力した文書画像に基づいて、PostScript文書およびHTML文書を生成する文書画像処理サーバ105bと、文書画像処理サーバ105a，105b（以下これらを「文書画像処理サーバ105」と記述する）で生成した文書を入力して、データベースに格納し、クライアント107からの検索要求に応じて、該当する文書を検索して出力する検索サーバ108と、から構成されている。なお、109は、LAN等のネットワークを示している。

【0016】図2は、図1に示した文書画像認識装置の概念構成図である。実施の形態1の文書画像認識装置は、大きく文書画像入力部200，前処理部201，情報抽出部204，文書生成部210およびデータベース部213から構成される。なお、文書画像入力部200，前処理部201，情報抽出部204および文書生成部210は、図1に示した文書画像処理サーバ105に該当し、データベース部213は、図1に示した検索サーバ108に該当する。以下に上記各部の構成について説明する。

【0017】（1）文書画像入力部

文書画像入力部200は、カラースキャナ100，モノクロスキャナ101，ネットワークスキャナ102およびディジタル複合機106で生成した文書画像やファックスモデム104で受信した文書画像を入力するものである。また、文書画像入力部200は、ワードプロセッサ等のアプリケーションプログラムで作成した文書ファイルを入力することもできる。

【0018】（2）前処理部

前処理部201は、文書画像入力部200を介して文書画像を入力し、入力した文書画像から孤立点ノイズを除去するノイズ除去部202と、入力した文書画像が傾いているような場合に、傾きを補正するスキュー補正部203とを有している。この前処理部201は、文書画像についてノイズ除去および傾き補正を行うことにより、後に行われる領域分割処理や文字認識処理において悪影響を及ぼす要因を除去するものである。なお、入力した文書画像がフルカラーであるような場合、領域分割処理等を容易に、かつ高速に行うことができるようにするため、フルカラーの文書画像を2値化する2値化処理を行うことにしても良い。

【0019】（3）情報抽出部

情報抽出部204は、前処理部201から文書画像を入力し、入力した文書画像から文字列を含む文字領域および／または図，表，写真等の画像を含む画像領域を認識して分割する処理を行うと共に、文書画像がいかなる段組種類、例えば、シングルコラム，マルチコラム，フリーコラムのいずれで構成されているかを識別する処理およびセンタリング領域を検知する処理等のレイアウト情報抽出処理を行う領域分割・情報抽出部205と、領域分割・情報抽出部205で分割した画像領域に表が含まれている場合に、表の枠と罫線の構造を抽出すると共に、枠内の文字領域を抽出する表処理部206と、領域分割・情報抽出部205および表処理部206で分割した文字領域の文字列について文字認識処理を行うOCR部207と、領域分割・情報抽出部205で分割した文字領域の文字列のフォントが強調系（中ゴシック等）や非強調系（細明朝等）のいずれであるかを識別する処理を行うフォント識別部208と、上記各部による処理の結果に基づいて、文書画像のレイアウトを解析すると共に、文書の論理的な構造を解析するレイアウト・論理解析部209と、を有している。

【0020】（4）文書生成部

文書生成部210は、OCR部207による文字認識処理の結果およびレイアウト・論理解析部209による解析処理の結果に基づいて、PostScript文書を生成するPostScript文書生成部211と、HTML文書を生成するHTML文書生成部212と、を有している。

【0021】（5）データベース（DB）

データベース213は、PostScript文書生成部211で生成されたPostScript文書を格納するPostScript文書DB214と、HTML文書生成部212で生成されたHTML文書を格納するHTML文書DB215とを有している。

【0022】次に、上述した構成を有する文書画像認識装置の動作について、詳細に説明する。図3は、文書画像認識装置の動作手順を示すフローチャートである。

【0023】文書画像入力部200は、カラースキャナ100，モノクロスキャナ101，ネットワークスキャナ102およびディジタル複合機106で生成した文書画像、並びにファックスモデム104で受信した文書画像を入力する（S301）。また、クライアント107から文書画像やワードプロセッサ等のアプリケーションプログラムで作成した文書ファイルを入力することもできる。

【0024】図4は、文書画像入力部200を介して入力した文書画像を画面表示した様子の一例を示す説明図である。図4において、400は文書画像認識処理の制御画面を、401は文書画像入力部200を介して入力した文書画像を表示する表示画面をそれぞれ示している。

【0025】図4に示す制御画面400では、以下に説明する各種処理の実行の指定および処理の詳細な設定を行うことができると共に、PostScript文書およびHTML文書の両方またはいずれか一方の生成を指定することができる。なお、図4においては、文書画像を入力し、各種処理の実行を指定することができる制御画面400を示したが、予め設定した条件に基づいて、文書の生成・格納まで自動的に実行できるようにすることもできる。

【0026】続いて、ノイズ除去部202は、文書画像入力部200を介して文書画像を入力し、入力した文書画像から孤立点ノイズを除去する（S302）。また、入力した文書画像が傾いているような場合、スキュー補正部203は、文書画像の傾きを補正する（S303）。

【0027】領域分割・情報抽出部205は、ノイズ除去部202およびスキュー補正部203からなる前処理部201から文書画像を入力し、入力した文書画像から文字列を含む文字領域および／または図，表，写真等の画像を含む画像領域を認識して分割する処理を行う（S304）。分割した各領域には、文字領域か、画像領域か、画像領域の場合にはさらに図，表，写真か等の領域の種類および領域の位置が属性情報として付与される。

【0028】領域分割処理を行った後、領域分割・情報抽出処理部205は、文書画像がいかなる段組種類、例えば、シングルコラム，マルチコラム，フリーコラムのいずれで構成されているかを識別する処理およびセンタリング領域の検知処理等のレイアウト情報を抽出する処理を行う（S305）。

【0029】具体的には、文字領域同士の間の距離（空

白部分）および罫線を検出し、検出した距離および罫線の本数と共に、領域分割処理で分割した文字領域の位置に関する属性情報に基づいて、段組種類の判定を行う。

【００３０】その後、表処理部２０６は、領域分割・情報抽出部２０５で分割した各領域の属性情報に基づいて、表を含む領域が存在するか否かを判定する（Ｓ３０６）。この判定は、上記のように装置側で自動的に行うことにしても良いし、ユーザが指定しても良い。表が含まれていない場合には、ステップＳ３０８に進む。

【００３１】一方、表が含まれている場合、表処理部２０６は、表の枠と罫線の構造を抽出すると共に、枠内の文字領域を抽出する（Ｓ３０７）。このように、表の中から文字領域を抽出することにより、次のＯＣＲ部２０７において、表中の文字認識処理を行うことができる。

【００３２】そして、ＯＣＲ部２０７は、領域分割・情報抽出部２０５および表処理部２０６で分割した文字領域の文字列について文字認識処理を行う（Ｓ３０８）。すなわち、ＯＣＲ部２０７は、文字領域について、行切り出しおよび文字切り出し処理を行い、文字切り出した個々の文字パターンについて文字認識処理を行う。加えて、ＯＣＲ部２０７は、文字認識結果に対して、言語処理による誤り補正を行う。

【００３３】フォント識別部２０８は、領域分割・情報抽出部２０５で分割した文字領域の文字列について、行単位でフォントが強調系（中ゴシック等）や非強調系（細明朝等）のいずれであるかを識別する処理を行う（Ｓ３０９）。具体的には、例えば、黒画素密度やランレングスの分布等に基づいて、フォントの特徴を識別する。

【００３４】続いて、レイアウト・論理解析部２０９は、上記各部による処理の結果に基づいて、文書画像のレイアウトを解析する（Ｓ３１０）。ここで行われるレイアウトの解析処理には、例えば、タイトル部，小見出し部，キャプション，ヘッダ・フッタ部の検出処理が含まれる。

【００３５】ここで、タイトルは、一般的に本文の文字とはサイズや行ピッチが異なり、また、存在する位置も本文とは若干離れていることから、領域分割・情報抽出部２０５で付与した領域の位置に関する属性情報や、フォント識別部２０８による識別結果を用いて、タイトル部を検出することができる。

【００３６】小見出しは、本文の文字と文字サイズがほぼ等しい場合も多く、本文に近接した場所に位置することから、本文と同一の領域に存在していることも多い。そこで、各文字領域の先頭行の文字サイズまたはフォントが、同一の文字領域中の他の文字のものと異なる場合に、先頭行を小見出し行と判定する。

【００３７】また、キャプションは、図，表，写真等の画像に付与されたものであり、一般的に画像領域の近傍で、本文とは離れた位置に存在すること、さらには、「図８」等の図や表等を指し示す語が含まれていることから、これらの条件を満たす行をキャプションと判定する。そして、キャプションと対応する図との間にリンクを生成する。

【００３８】さらに、ヘッダ・フッタは、文書画像の上下に存在するため、該当する行をヘッダ・フッタとして検出する。具体的には、例えば、領域識別・情報抽出部でシングルコラムと識別された場合、センタリング行より上の行がヘッダ部と判定することができる。この際、ヘッダの行のレベルについても識別する。また、マルチコラムと識別された場合、段に属しない上下の行がヘッダ・フッタ部となる。

【００３９】その後、レイアウト・論理解析部２０９は、文書の論理的な構造を解析する（Ｓ３１１）。具体的には、図，表，写真等に付与されたキャプション中の「図８」等の語と同一の語を本文中から見つけ出し、これらの文字間にリンクを生成する処理を行う。この処理においては、ＯＣＲ部２０７による文字認識の結果を用いるため、各文字パターンの候補文字の全てを使用することによって、ＯＣＲエラーに対処することができる。

【００４０】そして、文書生成部２１０は、上述した情報抽出部２０４の各部で行った処理の結果に基づいて、文書を生成する処理を行う（Ｓ３１２）。すなわち、ＰｏｓｔＳｃｒｉｐｔ文書生成部２１１は、情報抽出部２０４で行った文字認識やレイアウト解析等の結果に基づいて、ＰｏｓｔＳｃｒｉｐｔで各ページを表現した文書を生成する。なお、ＯＣＲエラーをカバーするため、文字認識確信度の低い文字については、文字認識の結果ではなく、元の画像を用いることができる。

【００４１】また、ＨＴＭＬ文書生成部２１２は、情報抽出部２０４で行った文字認識やレイアウト解析等の結果に基づいて、タイトル，段落等を示すタグを各文字列に付与し、ＨＴＭＬ文書を生成する。

【００４２】その後、ＰｏｓｔＳｃｒｉｐｔ文書はＰｏｓｔＳｃｒｉｐｔ文書ＤＢ２１４に、ＨＴＭＬ文書はＨＴＭＬ文書ＤＢ２１５にそれぞれ格納される（Ｓ３１３）。なお、同一の文書画像から生成したＰｏｓｔＳｃｒｉｐｔ文書とＨＴＭＬ文書の両方を登録する場合には、一方から他方を呼び出すことができるように、相互に関連づけを行うことにしても良い。

【００４３】なお、ＰｏｓｔＳｃｒｉｐｔ文書ＤＢ２１４およびＨＴＭＬ文書ＤＢ２１５にそれぞれ格納されたＰｏｓｔＳｃｒｉｐｔ文書およびＨＴＭＬ文書は、検索処理を行うことによって各クライアント１０７で画面表示することができる。すなわち、クライアント１０７からの検索要求に応じて、検索サーバ１０８がＰｏｓｔＳｃｒｉｐｔ文書ＤＢ２１４およびＨＴＭＬ文書ＤＢ２１５から該当する文書を検索して出力し、クライアント１０７は、検索サーバ１０８から検索結果を入力し、検索要求に該当する文書を画面表示する。

【００４４】図５は、ＰｏｓｔＳｃｒｉｐｔ文書を画面表示した様子の一例を示す説明図である。図５に示すように、図４に示した文書画像に基づいてＰｏｓｔＳｃｒｉｐｔで表現した文書を生成することにより、元の紙文書と同一のレイアウトの文書を容易に生成することができる。すなわち、紙文書を再現した文書を生成することにより、文書作成者の意図をできるだけ正確に伝えることができるような、流通および閲覧に適した文書を得ることができる。なお、ＰｏｓｔＳｃｒｉｐｔ文書を画面表示した様子は元の紙文書とほぼ同一であるが、内部情報は保持されているため、検索を行ったり、再利用したりすることができる。

【００４５】また、図６は、ＨＴＭＬ文書を画面表示した様子の一例を示す説明図である。図６は、図と図番号のハイパーテキスト化を行ったものであり、例えば、本文中の「図９」をマウス等でクリックすると、「図９」に該当する図が画面表示される。このように、紙文書が持つレイアウトにとらわれることなく、ハイパーテキスト化することにより、紙文書の内容を優先したコンピュータならではの形を持った文書を生成することができる。

【００４６】このように、実施の形態１の文書画像認識装置によれば、文書画像から文字コードを抽出するという単なる文字認識処理を行うだけではなく、文書画像の持つ様々な情報を抽出して利用するため、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を生成することができる。

【００４７】なお、図１においては、ネットワーク１０９を介したシステムとして実施の形態１の文書画像認識装置の構成を説明したが、図２に示す機能を１台のコンピュータに持たせることにより、スタンドアローンの形態で文書画像認識装置を構成することもできる。

【００４８】また、実施の形態１の文書画像認識装置では、上述したＰｏｓｔＳｃｒｉｐｔ文書やＨＴＭＬ文書を生成することにしたが、これらに限定するものではなく、必要に応じて他の形式の文書を生成することもできる。

【００４９】〔実施の形態２〕実施の形態２の文書画像認識装置は、実施の形態１で説明したようにして生成したＰｏｓｔＳｃｒｉｐｔ文書やＨＴＭＬ文書を効率良く検索することができるようにしたものである。具体的には、ＯＣＲ部２０７で文字認識した文字列から所定の文字列を抽出し、抽出した文字列を対応するＰｏｓｔＳｃｒｉｐｔ文書やＨＴＭＬ文書に関連づけておき、該当する文字列を検索することにより、関連づけされたＰｏｓｔＳｃｒｉｐｔ文書やＨＴＭＬ文書を検索結果として出力できるようにするものである。以下では、この文字列のことをキーテキストと定義することにする。

【００５０】図７は、実施の形態２の文書画像認識装置の概念構成図である。図７において、実施の形態１で説明した図２と同一の構成については同一の符号を付すことにし、それらの詳細な説明については省略する。

【００５１】実施の形態２の文書画像認識装置は、図７に示すように、ＯＣＲ部２０７で文字認識した文字列から上述したキーテキストを抽出するキーテキスト抽出部７００と、キーテキスト抽出部７００で抽出したキーテキストを入力し、キーテキストＤＢ７０２に登録するキーテキスト登録部７０１と、検索要求を入力し、キーテキストＤＢ７０２に登録されたキーテキストを検索して、該当するキーテキストに関連づけられたＰｏｓｔＳｃｒｉｐｔ文書またはＨＴＭＬ文書を検索結果として出力する検索処理部７０３と、を備えている。

【００５２】なお、ＰｏｓｔＳｃｒｉｐｔ文書ＤＢ２１４，ＨＴＭＬ文書ＤＢ２１５およびキーテキストＤＢ７０２は、検索処理部７０３に設けられる。この検索処理部７０３は、図１における検索サーバ１０８に該当し、クライアント１０７からの検索要求に基づいて、上記検索処理を行う。また、実施の形態２の文書画像認識装置をスタンドアローンの形態で構成した場合には、直接検索要求を入力して検索処理を行う。

【００５３】上記キーテキスト抽出部２０７で抽出するキーテキストとしては、文書を端的に表した文字列，例えば、文書全体，章，節のタイトルや、ヘッダ・フッタ等の書誌的事項，文書の要約文等が考えられる。また、文書中の図等を基準として、図のキャプションを構成する文字列や、図番を含むセンテンス，このセンテンスを含むパラグラフおよびページ単位の文字列をキーテキストとして抽出しても良い。なお、上記キーテキストを抽出するには、文書画像のレイアウトを解析する必要があることから、キーテキスト抽出部７００は、レイアウト・論理解析部２０９による解析結果を用いて、キーテキストの抽出処理を行うようにしても良い。

【００５４】また、キーテキスト抽出部７０１は、上記キーテキスト抽出部７００で抽出したキーテキストを入力し、入力したキーテキストを文書生成部２１０で生成したＰｏｓｔＳｃｒｉｐｔ文書やＨＴＭＬ文書に関連づけ、キーテキストＤＢ７０２に格納する。

【００５５】さらに、検索処理部７０３は、検索要求を入力すると共に、検索結果を出力する入出力部７０４と、入出力部７０４から検索要求を入力し、キーテキストＤＢ７０２から該当するキーテキストを検索する検索エンジン７０５とを有している。具体的に、入出力部７０４は、検索要求を入力して検索エンジン７０５に検索要求を出力する。検索エンジン７０５は、入出力部７０４から検索要求を入力し、キーテキストＤＢ７０２から該当するキーテキストを検索し、該当するキーテキストを入出力部７０４に出力する。入出力部７０４は、入力したキーテキストに関連づけられたＰｏｓｔＳｃｒｉｐｔ文書やＨＴＭＬ文書を検索結果として出力する。

【0056】このように、実施の形態2の文書画像認識装置によれば、文書画像中の文字列をキーテキストとして文書の検索を行うことにしたため、検索要求に対して、最も適切な検索結果を得ることができると共に、高速な検索処理を実現することができる。

【0057】なお、上述した実施の形態2においては、キーテキストを用いて文書の検索を用いることにしたが、キーテキストに代え、OCR部207で文字認識した結果全てを用いて全文検索を行うようにすることもできる。図8は、実施の形態2の文書画像認識装置の変形例を示す概念構成図である。

【0058】図8に示すように、上述したキーテキスト抽出部700，キーテキスト登録部701およびキーテキストDB702に代えて、テキスト登録部800および全文検索用テキストDB801を設け、テキスト登録部800がOCR部207から文字認識結果、即ち文書画像中のテキストの全文を入力し、入力したテキストを文書生成部210で生成したPostScript文書やHTML文書に関連づけし、全文検索用テキストDB801に登録する。その結果、全文検索用テキストDB801に登録されたテキストを用いて、PostScript文書やHTML文書を検索することができる。この場合は、テキストを全文検索用テキストDB801に登録するため、検索の度に各ファイルのオープン・クローズという処理が不要となるため、検索処理の高速化を図ることができる。

【0059】〔実施の形態3〕続いて、実施の形態3の文書画像認識装置について説明する。実施の形態3の文書画像認識装置は、実施の形態1のものと同様に、流通および閲覧を考慮した文書と、利用を考慮した文書を生成することができるようにしたものである。

【0060】実施の形態3の文書画像認識装置における流通および閲覧を考慮した文書とは、オリジナルの紙文書を読み取って生成した文書画像であり、ここではイメージデータと定義することにする。また、利用を考慮した文書とは、文書画像中の文字列について文字認識を行った結果であり、ここではテキストデータと定義することにする。

【0061】図9は、実施の形態3の文書画像認識装置の概念構成図である。なお、図9において、実施の形態1で説明した構成と同一の構成については同一の符号を付し、これらの詳細な説明は省略する。

【0062】実施の形態3の文書画像認識装置は，大きく文書画像入力部200，前処理部201，情報抽出部204，登録部900およびデータベース部213から構成される。なお，文書画像入力部200，前処理部201，情報抽出部204および登録部900は，図1に示した文書画像処理サーバ105に該当し、データベース部903は，図1に示した検索サーバ108に該当する。

【0063】文書画像入力部200および前処理部201については実施の形態1で説明した通りである。また、情報抽出部204においては、PostScript文書やHTML文書を生成しないため、フォント識別部208およびレイアウト・論理解析部209が省略されている以外は実施の形態1で説明した通りである。

【0064】実施の形態1の文書画像認識装置と異なるのは、イメージDB登録部901およびテキストDB登録部902を備えた登録部900と、イメージDB904およびテキストDB905を備えたデータベース部903である。

【0065】イメージDB登録部901は，前処理部201からノイズ除去およびスキュー補正を行った後のイメージデータを入力し、入力したイメージデータをデータベース部903のイメージDB904に登録する処理を行い、テキストDB登録部902は、OCR部207で文字認識処理を行った結果であるテキストデータをデータベース部903のテキストDB905に登録する処理を行う。なお、同一文書のイメージデータとテキストデータの両方を登録する場合には、一方から他方を呼び出すことができるように、相互に関連づけを行うことにしても良い。

【0066】なお、イメージDB904およびテキストDB905にそれぞれ格納されたイメージデータおよびテキストデータは、検索処理を行うことによって各クライアント107で画面表示することができる。すなわち、クライアント107からの検索要求に応じて、検索サーバ108がイメージDB904およびテキストDB905から該当する文書を検索して出力し、クライアント107は、検索サーバ108から検索結果を入力し、検索要求に該当する文書を画面表示する。

【0067】このように、実施の形態3の文書画像認識装置によれば、イメージデータをデータベースに登録するため、紙文書そのものをコンピュータ上で画面表示することができ、また、テキストデータをデータベースに登録するため、紙文書中の情報を容易に利用することができる。

【0068】〔実施の形態4〕実施の形態4の文書画像認識装置は、実施の形態3で説明したようにして生成したイメージデータやテキストデータを、実施の形態2で説明したキーテキストを用いて効率良く検索することができるようにしたものである。

【0069】図10は、実施の形態4の文書画像認識装置の概念構成図である。図10において、実施の形態1で説明した図2および実施の形態2で説明した図7と同一の構成については同一の符号を付すことにし、それらの詳細な説明については省略する。

【0070】実施の形態4の文書画像認識装置は、図10に示すように、文書画像のレイアウトを解析し、OCR部207で文字認識した文字列から上述したキーテキ

ストを抽出するキーテキスト抽出部700と、キーテキスト抽出部700で抽出したキーテキストを入力し、キーテキストDB702に登録するキーテキスト登録部701と、直接に検索要求を入力し、キーテキストDB702に登録されたキーテキストを検索して、該当するキーテキストに関連づけられたイメージデータやテキストデータを検索結果として出力する検索処理部703と、を備えている。

【0071】上記キーテキスト抽出部207で抽出するキーテキストとしては、文書を端的に表した文字列，例えば、文書全体，章，節のタイトルや、ヘッダ・フッタ等の書誌的事項、文書の要約文等が考えられる。また、文書中の図等を基準として、図のキャプションを構成する文字列や、図番を含むセンテンス，このセンテンスを含むパラグラフおよびページ単位の文字列をキーテキストとして抽出しても良い。

【0072】また、キーテキスト抽出部701は、上記キーテキスト抽出部700で抽出したキーテキストを入力し、対応するイメージデータやテキストデータに関連づけ、キーテキストDB702に格納する。

【0073】さらに、検索処理部703は、検索要求を入力すると共に、検索結果を出力する入出力部704と、入出力部704から検索要求を入力し、キーテキストDB702から該当するキーテキストを検索する検索エンジン705とを有している。具体的に、入出力部704は、検索要求を入力して検索エンジン705に検索要求を出力する。検索エンジン705は、入出力部704から検索要求を入力し、キーテキストDB702から該当するキーテキストを検索し、該当するキーテキストを入出力部704に出力する。入出力部704は、入力したキーテキストに関連づけられたイメージデータやテキストデータを検索結果として出力する。

【0074】上記検索処理部703は、図1における検索サーバ108に該当し、クライアント107からの検索要求に基づいて、検索処理を行う。また、実施の形態2の文書画像認識装置をスタンドアローンの形態で構成した場合には、直接検索要求を入力して検索処理を行う。

【0075】このように、実施の形態4の文書画像認識装置によれば、文書画像中の文字列をキーテキストとして文書の検索を行うことにしたため、検索要求に対して、最も適切な検索結果を得ることができると共に、高速な検索処理を実現することができる。

【0076】なお、上述した実施の形態4においては、キーテキストを用いて文書の検索を用いることにしたが、キーテキストに代え、OCR部207で文字認識した結果全てを用いて全文検索を行うようにすることもできる。図11は、実施の形態4の文書画像認識装置の変形例を示す概念構成図である。

【0077】図11に示すように、テキスト登録部800および全文検索用テキストDB801を設け、テキスト登録部800がOCR部207から文字認識結果、即ち文書画像中のテキストの全文を入力し、対応するイメージデータやテキストデータに関連づけし、全文検索用テキストDB801に登録する。その結果、全文検索用テキストDB801に登録されたテキストを用いて、イメージデータやテキストデータを検索することができる。この場合は、テキストを全文検索用テキストDB801に登録するため、検索の度に各ファイルのオープン・クローズという処理が不要となるため、検索処理の高速化を図ることができる。

【0078】なお、上述した実施の形態1～4の文書画像認識装置は、それぞれ任意に組み合わせて1つの文書画像認識装置を構成することができる。

【0079】さらに、上述した実施の形態1～4の文書画像認識装置としてコンピュータを機能させるプログラムを作成し、これらをハードディスク，フロッピーディスク，CD－ROM，MO，DVD等のコンピュータ読み取り可能な記録媒体に記録して、記録媒体を介してプログラムを配布することができる。そして，記録媒体に記録されたプログラムをコンピュータで読み出して実行することにより、上述した文書画像認識装置を実現することができる。

【0080】

【発明の効果】以上説明したように、本発明の文書画像認識装置（請求項1）によれば、紙文書を光学的に読み取ることによって生成した文書画像を入力する入力手段と、入力手段を介して入力した文書画像から文字列を含む文字領域および／または図，表，写真等の画像を含む画像領域を認識して抽出する領域抽出手段と、領域認識手段で抽出した文字領域の文字列について文字認識処理を行う文字認識手段と、領域抽出手段の抽出結果に基づいて、文書画像のレイアウトを解析し、レイアウト情報を抽出するレイアウト情報抽出手段と、文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、ページ記述言語を用いた第1の文書を生成する第1の文書生成手段と、文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、構造化記述言語を用いた第2の文書を生成する第2の文書生成手段と、第1および第2の文書生成手段で生成した第1および第2の文書をそれぞれ格納する格納手段と、を備えたため、文書画像から文字コードを抽出するという単なる文字認識処理を行うだけではなく、文書画像の持つ様々な情報を抽出して利用することにより、紙文書の再現を優先した文書や紙文書の内容を重視した文書等，利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を生成することができる。

【0081】また、本発明の文書画像認識装置（請求項2）によれば，請求項1に記載の文書画像認識装置にお

いて、第１の文書は、ＰｏｓｔＳｃｒｉｐｔ形式またはＰＤＦ形式によって表現された文書であるため、文書作成者の意図をできるだけ正確に伝えることができるような、流通および閲覧に適した文書を得ることができる。

【００８２】また、本発明の文書画像認識装置（請求項３）によれば、請求項１に記載の文書画像認識装置において、第２の文書は、ＳＧＭＬ，ＨＴＭＬまたはＸＭＬによって表現された文書であるため、紙文書が持つレイアウトにとらわれることなく、紙文書の内容を優先したコンピュータならではの形を持った文書を生成することができる。

【００８３】さらに、本発明のコンピュータ読み取り可能な記録媒体（請求項４）によれば、請求項１～３のいずれか１つに記載の文書画像認識装置の各手段としてコンピュータを機能させるためのプログラムを記録したため、記録したプログラムをコンピュータ上で実行することにより、文書画像の持つ様々な情報を抽出して利用することにより、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を生成することができる。

【図面の簡単な説明】

【図１】実施の形態１の文書画像認識装置のシステム構成を示す構成図である。

【図２】図１に示す文書画像認識装置の概念構成図である。

【図３】実施の形態１の文書画像認識装置の動作手順を示すフローチャートである。

【図４】実施の形態１の文書画像認識装置において、文書画像入力部を介して入力した文書画像を画面表示した様子の一例を示す説明図である。

【図５】実施の形態１の文書画像認識装置において、生成したＰｏｓｔＳｃｒｉｐｔ文書を画面表示した様子の一例を示す説明図である。

【図６】実施の形態１の文書画像認識装置において、生成したＨＴＭＬ文書を画面表示した様子の一例を示す説明図である。

【図７】実施の形態２の画像認識装置の概念構成図である。

【図８】実施の形態２の文書画像認識装置の変形例を示す概念構成図である。

【図９】実施の形態３の文書画像認識装置の概念構成図である。

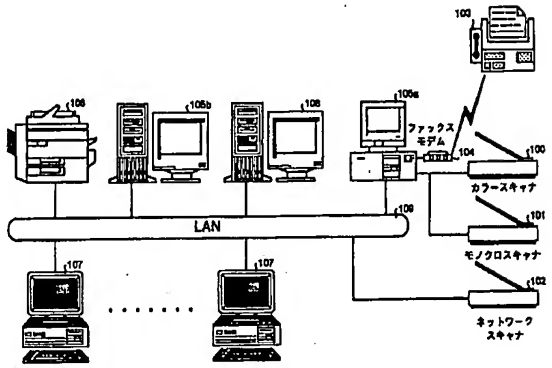
【図１０】実施の形態４の文書画像認識装置の概念構成図である。

【図１１】実施の形態４の文書画像認識装置の変形例を示す概念構成図である。

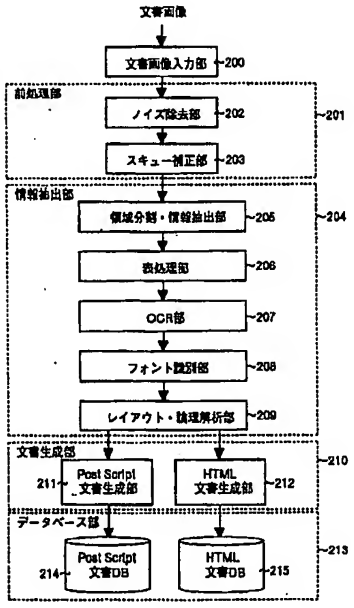
【符号の説明】

１００	カラースキャナ
１０１	モノクロスキャナ
１０２	ネットワークスキャナ
１０３	ファクシミリ装置
１０４	ファックスモデム
１０５（１０５ａ，１０５ｂ）	文書画像処理サーバ
１０６	ディジタル複合機
１０７	クライアント
１０８	検索サーバ１０８
１０９	ネットワーク
２００	文書画像入力部
２０１	前処理部
２０２	ノイズ除去部
２０３	スキュー補正部
２０４	情報抽出部
２０５	領域分割・情報抽出部
２０６	表処理部
２０７	ＯＣＲ部
２０８	フォント識別部
２０９	レイアウト・論理解析部
２１０	文書生成部
２１１	ＰｏｓｔＳｃｒｉｐｔ文書生成部
２１２	ＨＴＭＬ文書生成部
２１３	データベース部
２１４	ＰｏｓｔＳｃｒｉｐｔ文書ＤＢ
２１５	ＨＴＭＬ文書ＤＢ
７００	キーテキスト抽出部
７０１	キーテキスト登録部
７０２	キーテキストＤＢ
７０３	検索処理部
７０４	入出力部
７０５	検索エンジン
８００	テキスト登録部
８０１	全文検索用テキストＤＢ

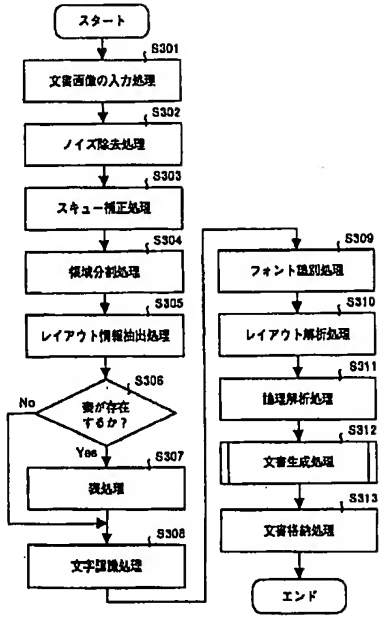
【図１】



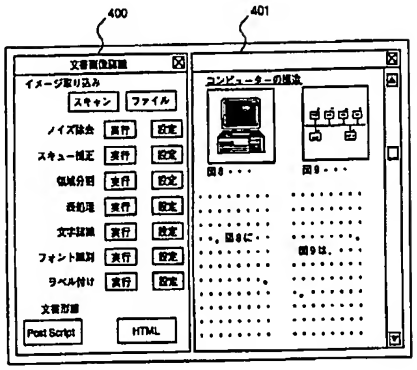
【図２】



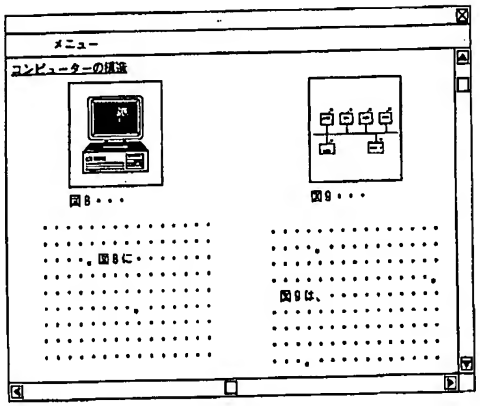
【図３】



【図4】


400
401
文書画像認識
イメージ取り込み
スキャン
ファイル
ノイズ除去
スキュー補正
領域分割
文字認識
フォント認識
ラベル付け
実行
設定
文書形態
Post Script
HTML
コンピューターの接続
図8・・・
図9・・・
図8に・・・
図9は、・・・

【図5】


メニュー
コンピューターの接続
図8・・・
図9・・・
図8に・・・
図9は、・・・

【図6】

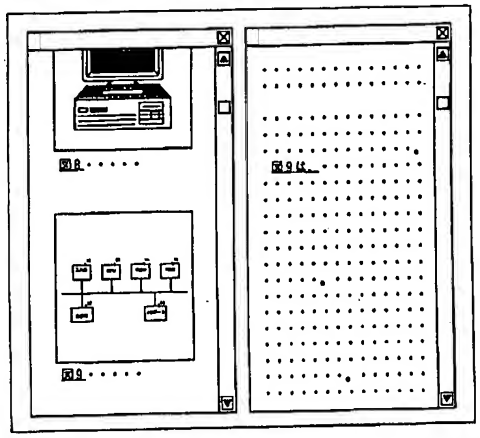
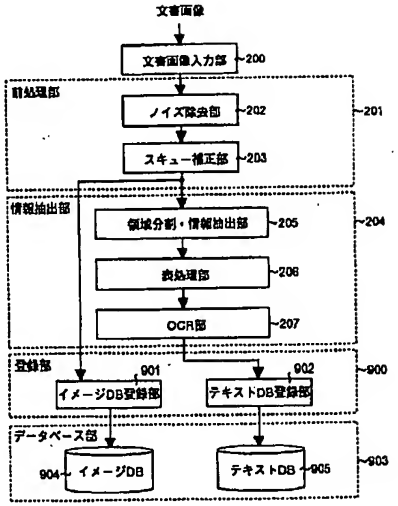
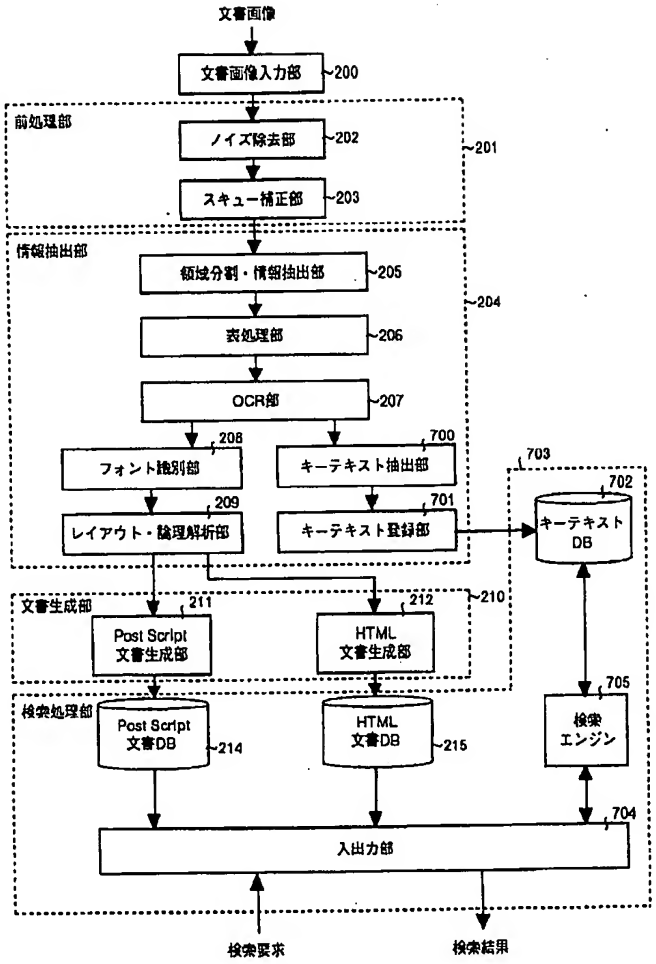


図8・・・・・
図9は、・・・
図9・・・・・

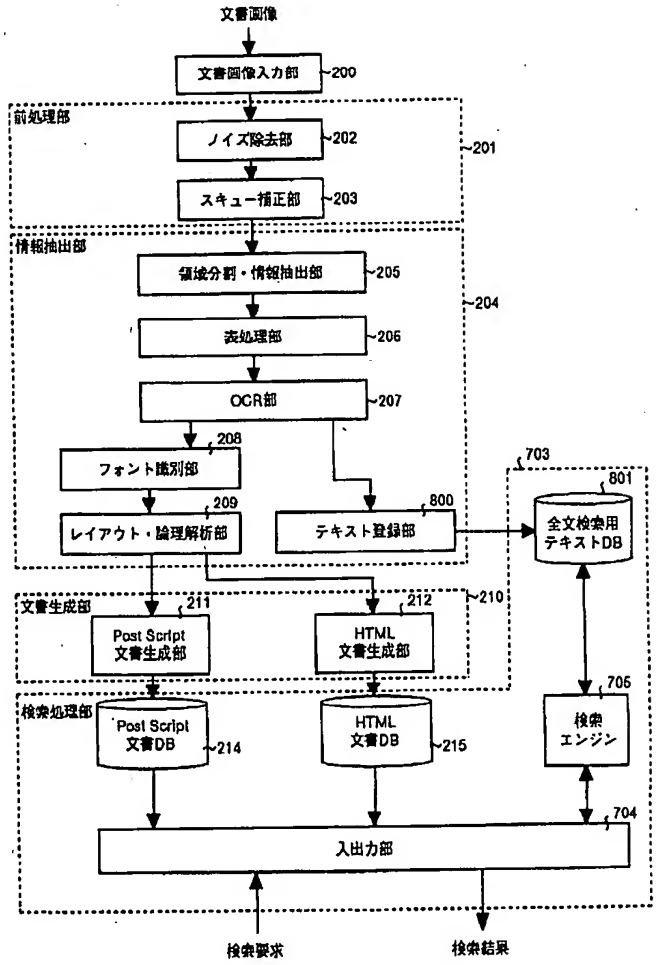
【図9】


文書画像
文書画像入力部
200
前処理部
ノイズ除去部
202
スキュー補正部
203
201
情報抽出部
領域分割・情報抽出部
205
表処理部
206
OCR部
207
204
登録部
901
902
イメージDB登録部
テキストDB登録部
900
データベース部
904
イメージDB
テキストDB
905
903

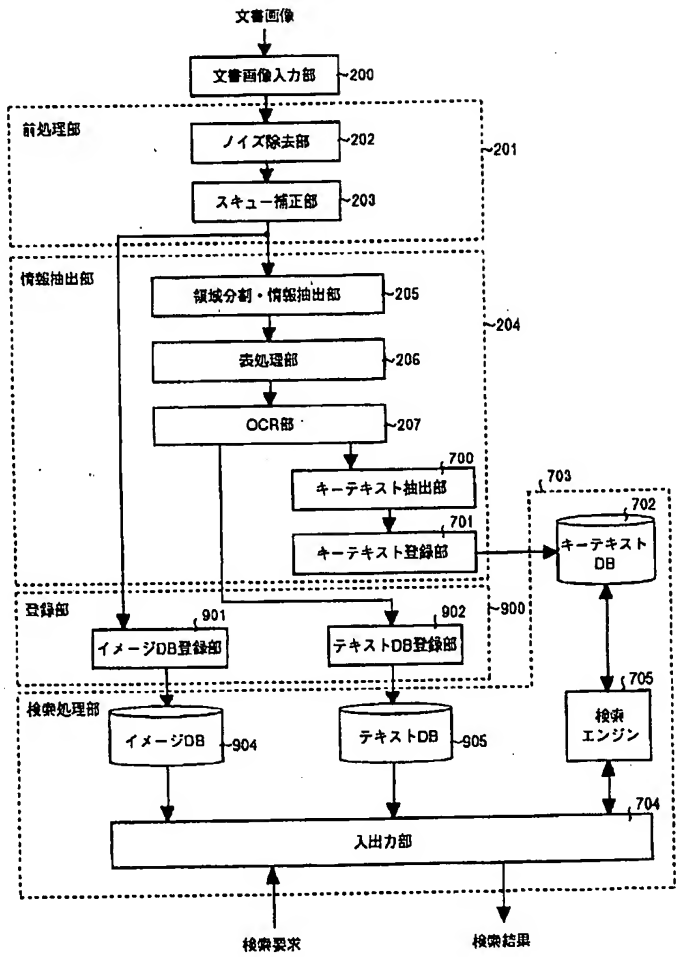
【図７】


文書画像
文書画像入力部
200
前処理部
ノイズ除去部
202
スキュー補正部
203
201
情報抽出部
領域分割・情報抽出部
205
表処理部
206
OCR部
207
204
フォント識別部
208
キーテキスト抽出部
700
レイアウト・論理解析部
209
キーテキスト登録部
701
703
キーテキスト
DB
702
文書生成部
211
Post Script
文書生成部
212
HTML
文書生成部
210
検索処理部
Post Script
文書DB
214
HTML
文書DB
215
検索
エンジン
705
704
入出力部
検索要求
検索結果

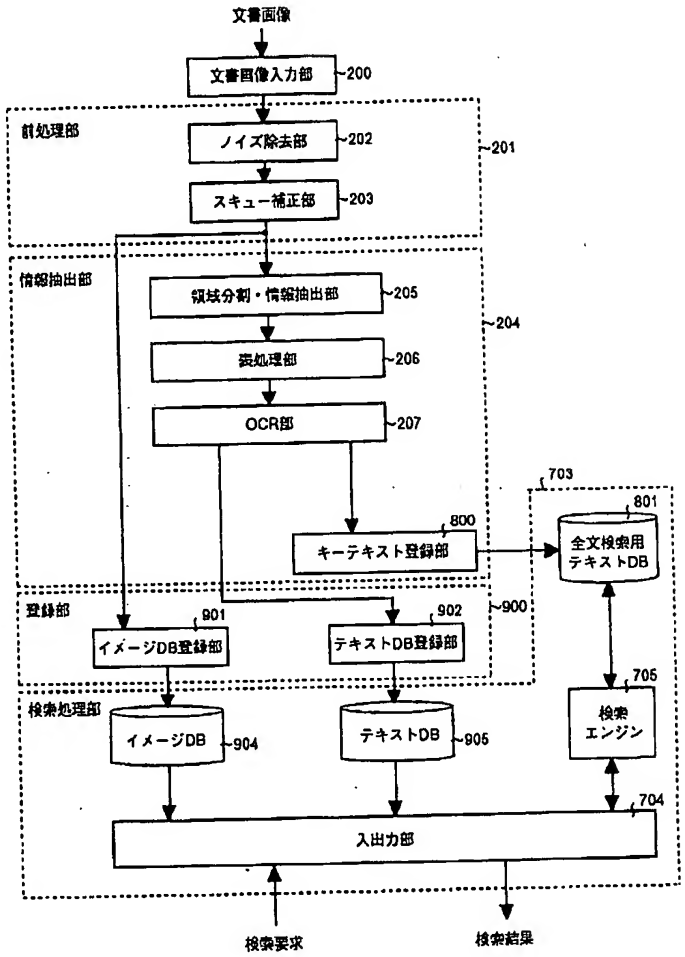
【図８】


文書画像
文書画像入力部
200
前処理部
ノイズ除去部
202
スキュー補正部
203
201
情報抽出部
領域分割・情報抽出部
205
表処理部
206
OCR部
207
204
フォント識別部
208
テキスト登録部
800
レイアウト・論理解析部
209
703
801
全文検索用
テキストDB
文書生成部
211
Post Script
文書生成部
212
HTML
文書生成部
210
検索処理部
Post Script
文書DB
214
HTML
文書DB
215
検索
エンジン
705
704
入出力部
検索要求
検索結果

【図10】



【図11】



フロントページの続き

(51) Int. Cl.[6]	識別記号	F I
H04N 1/40		H04N 1/40 F